Progress report 2022-01 – 2022-06

Modeling Cooperation

Paolo Bova, Jasmine Brazilek, Ben Harack, Vasily Kuznetsov,
Jonas Emanuel Müller, Tanja Rüegg, Miles Tidmarsh

Executive summary

Modeling Cooperation aims to gain better insight into the dynamics of competition for the development of transformative AI to look for opportunities to promote safety-enhancing cooperation among competitors. During the reporting period, we prepared the launch of our research software tool SPT Model which we built in collaboration with Associate Professor Robert Trager. The tool implements the Safety-Performance Tradeoff model created by Robert Trager, Paolo Bova, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe. It allows other researchers and decision-makers to explore how improvements in technical AI safety could affect the safety choices of competing AI developers.

We also created a test release of the software we are implementing in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop where participants explore AI governance scenarios. Furthermore, we advanced our revised and substantially extended technical report evaluating the Windfall Clause policy when applied to the model presented in the paper *Racing to the Precipice* through our internal reviewing process in preparation for uploading the new version to arXiv. Currently, Modeling Cooperation consists of seven team members corresponding to 2.2 FTE. While the full-time team members receive fixed wages of $30,000 annually, the employed part-time team member works for $15.625/hour on a contract basis.

Description

Modeling Cooperation aims to gain better insight into the dynamics of competition for the development of transformative AI (TAI) by building research software tools and conducting research to look for opportunities to promote safety-enhancing cooperation among competitors. We consider the mitigation and prevention of dangerous competition for TAI high-impact because such dynamics could strongly incentivize the competitors to underinvest in safety which in turn could lead to an increased risk of disaster.

Team

Modeling Cooperation entered 2022 with the support of a $83,000 grant from Jaan Tallinn as part of the Survival and Flourishing Fund. This allowed us to continue employing Paolo Bova as well as Jonas Emanuel Müller full-time and start paying our long-term volunteer Tanja Rüegg part-time. Ben Harack started to finance his part-time contributions through a personal grant he received from Survival and Flourishing Projects and, thus, is waiving his Modeling Cooperation wages. In combination with Jonas waiving 1,225 Swiss francs he earned as a freelancer outside of Modeling Cooperation during this reporting period, we can continue building up a small runway and, thus, increase the job security for our researchers.

Jasmine Brazilek has been continuing to support Modeling Cooperation part-time on a voluntary basis while Miles Tidmarsh and Vasily Kuznetsov have remained on hiatus. At the moment, our seven team members correspond to 2.2 FTE. Both full-time team members receive fixed wages of $30,000 annually and the employed part-time team member works for the corresponding hourly wage of $15.625 on a contract basis, which means she only gets paid when working and not when anything prevents her from doing so (e.g. sickness, vacations, COVID-19).

Strategy

At the beginning of the reporting period, we re-examined our planned work in light of our financial and human resources as well as our work to date and short-term goals. After completing the first release of our research software tool built in collaboration with Robert Trager and starting our collaboration with Shahar Avin to build software for his and his team's Intelligence Rising workshop during the second half of 2021, we decided to dedicate almost all of our resources to these two projects.

- Together with Robert Trager—an associate professor of international relations at UCLA and strategic modeling team lead at GovAI—we decided to continue working on our research software tool implementing an AI competition model with the goal of polishing and launching it during this reporting period. The tool will allow other researchers and decision-makers to explore how improvements in technical AI safety could affect the safety choices of competing AI developers.
- In our project with Shahar Avin, who is a senior research associate at CSER, we decided to continue implementing software for his and his team's Intelligence Rising workshop—a simulation game that allows participants to explore AI governance scenarios with other teams as they advance through an AI technology tree. The goal for this reporting period is to release the first version of the software and to conduct user field testing, i.e. letting the team of Intelligence Rising test the software built for facilitating the workshop in their actual working environment. Later on, the software will enable us to explore how access to different tools can influence the decision-making of participants, who may in the future hold key positions relevant to AI governance.

We assess these collaborations to be promising projects for Modeling Cooperation because they help to enable other researchers and policymakers to gain an intuitive understanding of AI competition models or AI development scenarios—a research direction our stakeholders repeatedly expressed interest in—and because we are exceptionally well-positioned to build user-friendly tools like web apps based on AI governance research results given our backgrounds and skills.

Additionally, we decided that we would spend some time building upon the following accomplishment we achieved during the previous reporting period: Revising and substantially extending our technical report in which we evaluated the Windfall Clause policy when applied to the model presented in the paper *Racing to the Precipice*. We aimed at working toward finalizing the incorporation of our new insights by advancing the report through our internal reviewing process. Once this process is complete, we will upload a revised version of our technical report to arXiv.

Accomplishments

- **Revised the Safety-Performance Tradeoff model we are implementing in our collaboration with Robert Trager and prepared the launch of our corresponding research software tool.** During the previous reporting period, we completed the first release of our research software tool displaying the AI competition model called the Safety-Performance Tradeoff model. While the model was initially created by Robert Trager, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe, team member Paolo Bova contributed substantially to the further development and analysis of the model during this project, and so was invited to become the second author of the corresponding working paper. During this reporting period, we focused on getting everything ready in preparation for the launch of our research software tool:
    - We continued to analyze the underlying model and found that the model had a pure strategy equilibrium in far fewer scenarios when laggards, not only leaders, were able to cause a disaster. This is important because pure strategies appear far more plausible than mixed strategies in contexts like international competition over AI between powerful actors and the lack of an equilibrium decreases the interpretability of the results. After discussing these new insights with our collaborators, we decided to revise the model by introducing the new parameter "laggard risk". This addition allows the users to intentionally explore the results under both assumptions and helps to identify a potentially important source of strategic uncertainty surrounding AI competition.
    - Announcing the research software tool to the relevant target audience is key to achieving our goal of allowing other researchers and decision-makers to explore how safety insights could affect the safety choices of competing AI developers. As a first step, we polished the tool with regard to its ease of use—increasing accessibility, clarity, and design. To do so, we substantially revised the UX based on user feedback, implemented a website tour that serves as an interactive walkthrough of the tool, added a description of the model as well as additional presets and new insights, implemented functionality to share specific scenarios via parameterized URLs, and implemented GDPR-compliant web analytics. As a second step, we wrote a blog post elaborating on the project which we shared with a number of

researchers from GovAI when we presented the new research software tool to them. After the presentation, we received valuable feedback regarding our planned blog post from the participants and Robert Trager expressed interest in launching the tool in the coming weeks.

- **Created a test release of the software we are implementing in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop.** This workshop allows participants to explore AI governance, and in particular AI competition, with other teams as they advance through an AI technology tree. Building upon the internal version we built during the second half of 2021, we focused on advancing three aspects of the software research tool in preparation for our test release: 1) Finishing the implementation of the core features allowing the game administrator to interact with the participants such as the login system and game administration section. 2) Setting up the infrastructure to run the application in a browser and synchronize the state of the game in real time between game administrators and participants. 3) Creating a first version of the tech tree through which the teams advance during the workshop. After completing these three work packages, we created a test release and held a presentation showing and explaining the functionality of the software to the game administrators.
- **Worked on finalizing the extension of our technical report *Safe Transformative AI via a Windfall Clause* about the evaluation of the Windfall Clause policy when applied to the model presented in the paper *Racing to the Precipice*.** Our technical report shows that in our model the Windfall Clause doesn't just encourage a safer race for TAI but that it is also often in the firms' best interests to pledge a significant share of windfall profits to socially good causes. While our report examined the public information setting for the design of a windfall clause, we succeeded in extending our findings to the private and no information settings present in previous works during the previous reporting period. Given these new results, we already revised and substantially extended our technical report and have now advanced it significantly through our internal reviewing process in preparation for uploading the new version to arXiv.

Roadblocks

- **Having new feature requests be introduced shortly before the planned release of the research software tool implementing the Safety-Performance Tradeoff model we built in collaboration with Robert Trager.** When continuing to analyze the Safety-Performance Tradeoff model, we gained new insights relevant to the results displayed in our research software tool. This led to the decision to revise the model shortly before the planned release. The additional effort resulted in not just postponing the launch of the aforementioned research software tool model but also in delaying the release of the first version of the software we implement to facilitate Shahar Avin and his team's Intelligence Rising workshop.

- **We could have benefitted from additional human resources which would have allowed us to reach our milestones quicker.** Currently, our team members cover a wide spectrum of different tasks ranging from software engineering over product management, design, and UX, to conducting research. Being spread so thin decreased our efficiency due to context switching and, thus, slowed down our progress. Being financially able to hire an additional team member and/or outsource specific tasks would have resolved this roadblock and allowed us to move our core work forward without slowdown.

Future goals

These goals are subject to change and contingent on funding as Modeling Cooperation aims to use its limited resources to investigate the most impactful research questions within AI governance. Thus, we regularly re-evaluate our planned work taking new research opportunities and feedback from our stakeholders into consideration.

Short-term goals

- **Launch our research software tool implementing the Safety-Performance Tradeoff model we built in collaboration with Robert Trager.** During this reporting period, we revised the Safety-Performance Tradeoff model and prepared the launch of our research software tool displaying the AI competition model created by Robert Trager, Paolo Bova, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe. Now it's time to get the word out. Before doing so, we plan to incorporate the feedback we received regarding our planned blog post after presenting the tool to other AI governance researchers and to apply some finishing touches. Once we are ready for the launch, we will announce the research software tool together with Robert Trager, among other channels, on the Effective Altruism forum to allow other researchers and decision-makers to explore how safety insights could affect the safety choices of competing AI developers.

- **Release the first version of the software we implemented in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop.** After creating a test release and holding a presentation showing and explaining the functionality of the software to the game administrators during this reporting period, we plan to finalize the first version of the software with the goal of user field testing, i.e. letting the team of Intelligence Rising test the software in their actual working environment when conducting a workshop. Before doing so, we have to finish the implementation of the tech tree through which participants advance to explore AI governance scenarios with other teams and migrate our database and our framework due to the renaming of the database we currently use.

Longer-term goals

- **Finalize the extension of our technical report *Safe Transformative AI via a Windfall Clause* and upload the new version of our report to arXiv.** In addition to evaluating a Windfall Clause, designed in a public information setting and applied to the model presented in the paper *Racing to the Precipice*, we managed to extend our findings to the private and no information settings present in previous works during this reporting period. Building upon this success, we were also able to examine when disclosure of information is helpful or harmful to a Windfall Clause. We plan to incorporate our new insights into our revised technical report which we then would like to publish on arXiv.
- **Further investigate the Windfall Clause policy when applied to a model with more parameters than the one presented in the paper *Racing to the Precipice*.** For analysis similar to the examination of the Windfall Clause policy, we think it is useful to start with a simple AI competition model to gain a basic but thorough understanding of the policy's most important dynamics. Once we author a paper about our initial analysis, we then plan to advance our research by applying the Windfall Clause policy to a more complex model which is more representative of real-world AI developments.
- **Expand our ongoing and start new collaborations with AI governance academics to build tools for other researchers and policymakers to help build intuitions for AI competition.** When discussing the impact of potential research directions, our stakeholders repeatedly expressed interest in projects which enable other researchers and policymakers to gain an intuitive understanding of AI competition models or AI development scenarios. Given the background and skills of Modeling Cooperation's team members, we would argue that we are

exceptionally well-positioned to build user-friendly tools like web apps based on AI governance research results. This view was strengthened during our ongoing collaboration with Robert Trager, who already expressed interest in continuing the project as well as significantly expanding its scope in the future. In the long run, we might want to work toward building an AI governance modeling platform that enables other researchers and policymakers to explore our own computational models and those of our collaborators.

- **Build upon the achievements Jonas accomplished during the project "Using Bayesian ML to find more interpretable solutions to AI race models":** Having been awarded a personal grant by Survival and Flourishing, Jonas worked on making numerical optimal solutions of an AI race model interpretable by humans using MIT's new Bayesian program learning ML system DreamCoder.
    - **Submit DreamCoder improvements implemented as part of the aforementioned project:** Jonas would like to contribute to the DreamCoder project by submitting his improvements to the researchers. To do so, he plans to document the new infrastructure and polish the bug fixes and extensions. We hope that these improvements help the researcher at MIT as well as other researchers outside of MIT to benefit from the DreamCoder ML system, which provides a safer alternative to current approaches to AGI.
    - **Refine the interpretable strategies gained as part of the aforementioned project and incorporate the solutions into our AI competition simulations:** Jonas would like to refine the results in terms of improving the accuracy as well as robustness and training DreamCoder to provide solutions for games with more dimensions. Afterward, he would like to use the resulting code from DreamCoder to make the agents in our existing AI competition simulation behave optimally.
- **Write a series of blog posts elaborating on Modeling Cooperation's approach, methods, and models.** One goal of our relaunched website is to allow anyone to stay up to date with our work. In addition to announcing our research results and research support software tools, we also want to share the insights we gained during our work for other researchers to benefit from. One example is our comparison of multiple analytical methods to solve dynamic AI competition games which could help other researchers to decide on a suitable approach without having to try out the various approaches themselves. Another example is our review of often technical and complex literature relevant to AI competition,

such as game theory and industrial organization, which we could write up tailored to other AI competition researchers.

- **Integrate statistical features into our Monte Carlo simulation tool.** We would like to integrate features such as null hypothesis significance testing, parameter sampling, and t-digest support to further increase our economists' productivity.